

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-096177

(43)Date of publication of application : 09.04.1999

(51)Int.Cl.

G06F 17/30

G06F 17/22

G06F 17/27

G06F 17/28

(21)Application number : 09-257364

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

(22)Date of filing : 22.09.1997

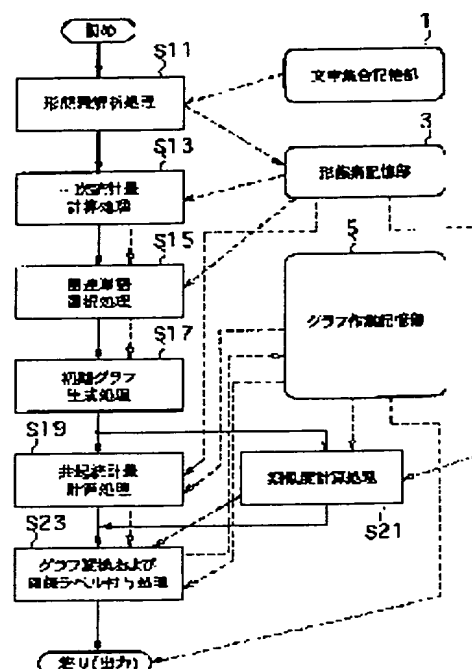
(72)Inventor : YUGAWA TAKASHI

(54) METHOD FOR GENERATING TERM DICTIONARY, AND STORAGE MEDIUM RECORDING TERM DICTIONARY GENERATION PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a term dictionary generating method capable of generating a term dictionary including information necessary for the processing of many documents over a wide object area by generating an ontology capable of recognizing various relation of words and a recording medium recording a term dictionary generation program.

SOLUTION: Respective words in a document are stored together with their positional information and a primary statistic value related to the inclusion of the same word is calculated (step S13). Relative words are selected based on the primary statistic value (step S15) and a graph linked with nodes of respective relative words is generated from the nodes of words expressing the object areas of the relative words (step S17). Then, a cooccurrence statistic value for the combination of two nodes of the graph is calculated (step S19) and similarity between two combined words is calculated (step S21). Then, an ontology is generated by converting the graph based on the cooccurrence statistic value and the similarity and annexing a relative label to the converted graph (step S23) to generate a term dictionary.



LEGAL STATUS

[Date of request for examination] 01.02.2001

[Date of sending the examiner's decision of rejection] 06.04.2004

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-96177

(43) 公開日 平成11年(1999) 4月9日

(51) IntCl.⁶

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/40

3 7 0 J

17/22

15/20

5 1 4 U

17/27

5 5 0 F

17/28

15/38

C

15/401

3 2 0 Z

審査請求 未請求 請求項の数 2 O L (全 8 頁) 最終頁に続く

(21) 出願番号

特願平9-257364

(71) 出願人 000004226

日本電信電話株式会社

東京都新宿区西新宿三丁目19番2号

(22) 出願日

平成9年(1997) 9月22日

(72) 発明者 湯川 高志

東京都新宿区西新宿三丁目19番2号 日本

電信電話株式会社内

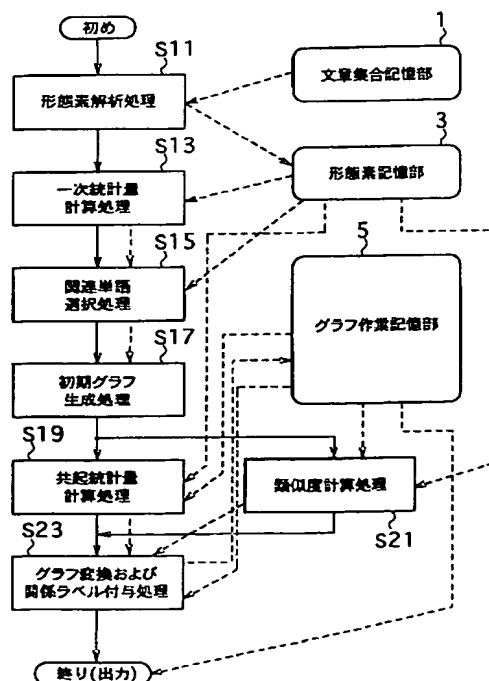
(74) 代理人 弁理士 三好 秀和 (外1名)

(54) 【発明の名称】 用語辞書生成方法および用語辞書生成プログラムを記録した記録媒体

(57) 【要約】

【課題】 単語の種々の関係を認定できるオントロジを生成して広い対象領域にわたる大量の文書の処理に必要とされる情報を含む用語辞書を生成し得る用語辞書生成方法および用語辞書生成プログラムを記録した記録媒体を提供する。

【解決手段】 文書の各単語の位置情報と共に格納し、同一単語が含まれることに関する一次統計量を計算し(ステップS13)、該一次統計量に基づいて関連単語を選択し(ステップS15)、関連単語の対象領域を表す単語のノードから各関連単語のノードにリンクを張ったグラフを生成し(ステップS17)、該グラフの各2ノードの組合せについて共起統計量を計算し(ステップS19)、各組合せの2つの単語の類似度を計算し(ステップS21)、共起統計量と類似度に基づきグラフを変換し、関係ラベルを付与し(ステップS23)、オントロジとして生成し、用語辞書を生成する。



【特許請求の範囲】

【請求項 1】 文書に用いられている単語の意味および使われ方を記憶した用語辞書を生成する用語辞書生成方法であって、

文書を読み込んで単語の列に分解し、該単語列の中の個々の単語を該単語の文書中の位置情報とともに格納し、前記単語列に含まれる単語について、該単語列に同一単語が含まれることに関する統計量を一次統計量として計算し、

この計算された各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、

この選択した関連単語の各々をノードとし、対象領域を代表的に表す単語のノードから前記関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、

この生成されたグラフのノードのあらゆる 2 つのノードの組合せについて、各組合せの 2 つの単語の前記位置情報に基づいて該 2 つの単語の同時出現についての統計量である共起統計量を計算し、

前記各組合せの 2 つのノードに対応する 2 つの単語の類似度を計算し、

前記共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与することを特徴とする用語辞書生成方法。

【請求項 2】 文書に用いられている単語の意味および使われ方を記憶した用語辞書を生成する用語辞書生成プログラムを記録した記録媒体であって、

文書を読み込んで単語の列に分解し、該単語列の中の個々の単語を該単語の文書中の位置情報とともに格納し、前記単語列に含まれる単語について、該単語列に同一単語が含まれることに関する統計量を一次統計量として計算し、

この計算された各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、

この選択した関連単語の各々をノードとし、対象領域を代表的に表す単語のノードから前記関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、

この生成されたグラフのノードのあらゆる 2 つのノードの組合せについて、各組合せの 2 つの単語の前記位置情報に基づいて該 2 つの単語の同時出現についての統計量である共起統計量を計算し、

前記各組合せの 2 つのノードに対応する 2 つの単語の類似度を計算し、

前記共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与することを特徴とする用語辞書生成プログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、文書に用いられて

いる単語の意味および使われ方を記憶した用語辞書を生成する用語辞書生成方法および用語辞書生成プログラムを記録した記録媒体に関し、特に平文テキスト形式で電子的に格納された文書をコンピュータが理解することによって処理を行う自然言語処理技術、情報検索技術および情報整理統合技術に有効な用語辞書生成方法および用語辞書生成プログラムを記録した記録媒体に関する。

【0002】

【従来の技術】 文書で用いられる単語の意味や使われ方を記述した辞書はオントロジと呼ばれ、平文で記述された文書をコンピュータで理解したり、理解に基づいて検索、分類または統合などの処理をする場合に必須となる。オントロジは、用語の意味や使われ方を、語と語の間の関係を通じて表している。このオントロジをコンピュータの記憶として格納したものを用語辞書と呼ぶことにする。用語辞書の表現形式としては、フレーム、一階述語、グラフ等が用いられるが、これらは本質的には同じである。

【0003】 文書をコンピュータに理解させ、理解に基づいて検索、分類または統合などの処理をさせる目的は、該コンピュータの利用者（以下、ユーザと呼ぶ）がインターネットや社内情報システム等に蓄積された大量の文書から、ユーザが必要としたり、あるいは興味を持つ情報を整理・統合された形態で得るためである。本発明が対象とするのは、蓄積された膨大な文書からユーザが興味を持つ対象領域に関連する文書が抽出され限定された後のコンピュータにおける処理である。ユーザは人間であり様々な対象に興味を持つため、取り得る対象領域は広い範囲にわたるが、ユーザがひとたび対象領域を限定すると、文書集合には該対象領域に関する事柄を記述した文書だけが含まれることになる。このことから、該コンピュータによる文書の理解、検索、分類、統合のための用語辞書は、対象領域が限定された後には、対象領域に関連のある語を含めばよいことになる。もちろん、対象領域そのものは広い範囲を取り得るため、どのような対象領域に対しても、該対象領域に関連する語を含む用語辞書は該文書集合が与えられた後には存在していなければならない。

【0004】 従来は、オントロジは人間が自己の言語に関する知識や、辞書や百科事典の記述などに基づいて手作業で構築し、該オントロジをコンピュータに入力することで用語辞書を生成していた。また、近年、語の文書上の出現頻度に基づいた類似度を計算し、相互に類似度が高い語の間に連想関係を認定することによって用語辞書を自動生成する方法が提案されている（参考文献：岩爪道昭、白神謙吾、武田英明、西田豊明、「インターネットからの情報収集・分類・統合化のためのオントロジ一獲得」、1996年度人工知能学会全国大会、18-03） 本来、オントロジにおける語の間の関係は連想関係だけではないため、該用語辞書生成方法で生成され

る用語辞書は、粗い構造のオントロジ（弱構造化オントロジ）を格納したものとなる。

【0005】

【発明が解決しようとする課題】人間の手作業によるオントロジの構築は、人間の判断速度の制約があるため、小規模な語の集合に対してさえ非常に時間がかかることになる。このことから、対象領域が限定されて文書集合が与えられた時点で、対象領域に関連する文書に含まれる用語の用語辞書を動的に生成することは、文書の規模が非常に小さくない限り不可能である。したがって必然的に、あらゆる文書に含まれると想定される語の集合に対する用語辞書を予め生成しておくことが必要となる。しかしながら、様々な対象領域のあらゆる文書に含まれると想定される語の数は非常に多く、また対象領域によって意味や使われ方の異なる単語も多く存在する。このため、生成に莫大な時間が必要とされる上、整合性が保持できない。このように、人間の手作業によるオントロジ構築では、インターネットや社内情報処理システム等に蓄積されているような広い対象領域にわたる大規模な文書テキストを処理するための用語辞書の生成は不可能である。

【0006】一方、上記の岩爪らの方法では、連想関連のみの認定による粗い構造のオントロジのため、該方法により生成された用語辞書は、文書の処理に必要とされる情報を十分には含んでいない。

【0007】本発明は、上記に鑑みてなされたもので、その目的とするところは、単語の種々の関係を認定できるオントロジを動的に生成して、広い対象領域にわたる大量の文書に対しても文書の処理に必要とされる情報を十分に含む用語辞書を生成し得る用語辞書生成方法および用語辞書生成プログラムを記録した記録媒体を提供することにある。

【0008】

【課題を解決するための手段】上記目的を達成するため、請求項1記載の本発明は、文書に用いられている単語の意味および使われ方を記憶した用語辞書を生成する用語辞書生成方法であって、文書を読み込んで単語の列に分解し、該単語列の中の個々の単語を該単語の文書中の位置情報とともに格納し、前記単語列に含まれる単語について、該単語列に同一単語が含まれることに関する統計量を一次統計量として計算し、この計算された各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、この選択した関連単語の各々をノードとし、対象領域を代表的に表す単語のノードから前記関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、この生成されたグラフのノードのあらゆる2つのノードの組合せについて、各組合せの2つの単語の前記位置情報に基づいて該2つの単語の同時出現についての統計量である共起統計量を計算し、前記各組合せの2つのノードに対応する2つの単語

の類似度を計算し、前記共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与することを要旨とする。

【0009】請求項1記載の本発明にあっては、文書を単語の列に分解し、該単語列の中の各単語をその位置情報とともに格納し、単語列に同一単語が含まれることに関する一次統計量を計算し、この各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、この関連単語の各々をノードとし、対象領域を表す単語のノードから関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、このグラフのノードのあらゆる2つのノードの各組合せの2つの単語の位置情報に基づいて該2つの単語の共起統計量を計算し、各組合せの2つのノードに対応する2つの単語の類似度を計算し、共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与して、用語辞書を生成しているため、文書の処理に必要とされる十分な情報を含んだ用語辞書を生成することができる。

【0010】また、請求項2記載の本発明は、文書に用いられている単語の意味および使われ方を記憶した用語辞書を生成する用語辞書生成プログラムを記録した記録媒体であって、文書を読み込んで単語の列に分解し、該単語列の中の個々の単語を該単語の文書中の位置情報とともに格納し、前記単語列に含まれる単語について、該単語列に同一単語が含まれることに関する統計量を一次統計量として計算し、この計算された各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、この選択した関連単語の各々をノードとし、対象領域を代表的に表す単語のノードから前記関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、この生成されたグラフのノードのあらゆる2つのノードの組合せについて、各組合せの2つの単語の前記位置情報に基づいて該2つの単語の同時出現についての統計量である共起統計量を計算し、前記各組合せの2つのノードに対応する2つの単語の類似度を計算し、前記共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与することを要旨とする。

【0011】請求項2記載の本発明にあっては、文書を単語の列に分解し、該単語列の中の各単語をその位置情報とともに格納し、単語列に同一単語が含まれることに関する一次統計量を計算し、この各単語の一次統計量に基づいて、対象領域に関連の深い単語を関連単語として選択し、この関連単語の各々をノードとし、対象領域を表す単語のノードから関連単語の各々のノードに対してそれぞれ有向リンクを張ったグラフを生成し、このグラフのノードのあらゆる2つのノードの各組合せの2つの単語の位置情報に基づいて該2つの単語の共起統計量を計算し、各組合せの2つのノードに対応する2つの単語

の類似度を計算し、共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与して、用語辞書を生成する用語辞書生成プログラムを記録媒体として記録しているため、該記録媒体を用いて、その流通性を高めることができる。

【0012】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態について説明する。

【0013】図1は、本発明の一実施形態に係る用語辞書生成方法の処理の流れおよび該処理を実施するのに必要な構成要素の一部を示した図である。なお、図1において、実線は処理の流れを示し、点線はデータの流れを示している。

【0014】図1を参照して、本実施形態の用語辞書生成方法について概略的に説明する。図1に示す用語辞書生成方法では、文書集合記憶部1に記憶されている文書を形態素解析処理によって単語の列に分解し（ステップS11）、この分解された単語の列に含まれる個々の単語をその文書内の位置情報とともに形態素記憶部3に記憶する。

【0015】次に、形態素記憶部3に記憶された単語の列において各単語の出現に関する統計量である一次統計量を一次統計量計算処理で計算する（ステップS13）。それから、この一次統計量に基づいてユーザが興味を持っている対象領域に関連の深い単語の集合を前記単語列から関連単語選択処理で選択する（ステップS15）。この単語の集合は次のステップS17の初期グラフ生成処理において対象領域を代表的に表す単語をルート・ノードとし、単語の集合の個々の単語をノードとして、ルート・ノードから個々のノードに向かって有向リンクを張った初期グラフを生成し、グラフ作業記憶部5に記憶される。

【0016】それから、このグラフ作業記憶部5に記憶された初期グラフに含まれる単語から2個のあらゆる単語の組合せについて、該2個の単語のペアが同一の文書や段落や文に同時に出現して含まれるかどうかに基づいた統計量である共起統計量を計算するとともに（ステップS19）、該単語のペアが意味的に類似しているかどうかについての類似度を計算する（ステップS21）。なお、この共起統計量計算処理と類似度計算処理は相互のデータ依存関係がないため、どちらを先に処理してもよく、また並行して処理してもよい。

【0017】共起統計量と類似度が計算されると、これらの値を利用する変換ルールとラベル付けルールに基づいてグラフ作業記憶部5に記憶されたグラフを変換し、リンクに関係ラベルを付与し、これによりグラフ作業記憶部5にオントロジを表すグラフが記憶される（ステップS23）。このオントロジを表すグラフは、グラフ中のノードが単語に対応し、リンクが単語と単語の関係に対応しているため、用語辞書となっているものである。

【0018】なお、オントロジを表すグラフをどのように出力するかは、本発明で規定するものではないが、グラフそのままの形式でグラフィック出力装置に出力することも可能であり、また一般にはグラフ作業記憶部5に記憶された状態そのまま、または単純な形式変換によって文書理解プログラムなどの他のプログラムに渡されて利用されることになる。

【0019】次に、図2以降の図面も参照して、図1に示す用語辞書生成方法について更に詳細に説明する。なお、この説明では、パーソナルコンピュータのSCSIカードに関する文書を対象として説明する。

【0020】最初に、パーソナルコンピュータのSCSIカードに関する記事が大量に与えられ、文書集合記憶部1に記憶されているものとする。なお、本実施形態では、ユーザが興味を持っている分野、すなわちパーソナルコンピュータのSCSIカードに関連する文書を収集する方法については特に規定しないが、これは、例えば電子会議室、ネットニュース、ワールドワイドウェブ等から収集することが考えられる。

【0021】処理が開始すると、まず形態素解析処理が行われる（ステップS11）。これは、文書集合記憶部1に記憶されている文書を単語の列に分解する。分解された単語列は、その所在情報とともに形態素記憶部3に記憶される。形態素記憶部3は、図2に示すようなテーブル構造を有する。

【0022】図2に示す形態素記憶部3のテーブル構造において、単語の欄には文書を分解した個々の単語が入り、文書IDは該単語が出現する文書を特定するための符号、段落IDは該単語が出現する段落を指定するための符号、文IDは該単語が出現する文を特定するための符号である。個々の符号の付与については本発明で規定されるものではないが、文書ID、段落ID、文IDの組によって文書、段落、文それぞれが一意に決定できるものでなければならない。すなわち、最も単純な付与法は、文書集合記憶部1に含まれるすべての文書を通じて、個々の文書に一意の文書IDを付与し、個々の段落に一意の段落IDを付与し、個々の文に一意の文IDを付与することである。文書が大量ならば、文の数は膨大であるから、文IDを整数として符号化するならば大きな値を持つことになる。文IDの最大値を少なくする付与法としては、文書集合記憶部1に含まれるすべての文書を通じて個々の文書に一意の文書IDを付与し、個々の段落に文書内で一意となるように段落IDを付与し、個々の文の段落内で一意となるような文IDを付与する方法が考えられる。本実施形態における図2では後者の方法で文書ID、段落ID、文IDを付与している。

【0023】なお、本実施形態では、個々の単語と該単語の所在情報をテーブルで表現して形態素記憶部3に格納しているが、本発明では該記憶をテーブル形式の表現として限定するものではない。単語と該単語の所在に関

する情報を関連づけて記憶できるならばどのような表現形式でもよく、他にリストや一階述語などが考えられる。また、所在情報に関しても、本実施形態では個々の単語について、該単語が出現する文書、段落、そして文を以って所在情報としているが、本発明での所在情報はこれらに限られるものではない。他に文字位置などが考えられる。

【0024】形態素解析処理の次に一次統計量処理を行う(ステップS13)。一次統計量計算処理では、形態素記憶部3に記憶された単語と該単語の所在情報に基づき、単語の出現に関する一次統計量を計算する。本実施形態において一次統計量計算は、図3に示すようなテーブルの生成と該テーブルに基づく計算によって行われる。図3のテーブルは、形態素記憶部3に記憶された単語に関して、同一単語が出現する回数を計数したものである。計数の方法については本発明で規定するものではないが、最初に図3のようなテーブルを生成して出現回数をすべて0にセットし、単語が出現するごとに該当する単語に対応して出現する回数を1だけインクリメントする方法が考えられる。該テーブルに基づいて、個々の単語に対し、該単語の出現回数を出現回数の合計値、すなわち文書集合記憶部1に記憶されているすべての文書に含まれる単語の全数で割算し、これを該単語の一次統計量とする。

【0025】なお、一次統計量は、文書集合記憶部1に含まれるすべての文書内での個々の単語の出現の重要度を示す統計量であれば、どのようなものでもよく、本実施形態で説明する統計量に限定されるものではない。また、一次統計量の計算法に関しても、本実施形態では出現回数に関するテーブルを生成した後合計によって除すことで計算したが、これは本発明で本質的に規定されるものではなく、一次統計量が得られるのであればどのような方法でもよい。他の方法として、最初に文書集合記憶部1に記憶されているすべての文書に含まれる単語の全数を計数しておき、単語が現れるたびに該単語に対する統計量に、1を単語全数で除した値を加算していく方法などが考えられる。

【0026】一次統計量計算処理が終了すると、該一次統計量計算処理で得られた一次統計量を用いて関連単語選択処理を行い(ステップS15)、ユーザが興味を持っている分野に関連の深い単語、すなわちこの場合パーソナルコンピュータのSCSIカードに関連の深い単語を1個以上選択する。該関連の深い単語の選択は一次統計量を用いて行われる。最も簡単なのは、一次統計量としてあるしきい値よりも大きい値を持つ単語を選択することであるが、本発明では選択の方法としてこれに限定するものではない。他に、一次統計量の大きい順に一定個数を選択する方法や、一次統計量の度数分布からしきい値を動的に決定して該閾値以上の一次統計量を持つ単語を選択する方法などが考えられる。更に、ここで得ら

れた個々の単語に関する一次統計量の他に、すでに広く一般に開示されている日常語や専門用語に関する統計量を利用し、該一次統計量を補正して単語選択に利用する方法なども考えられる。

【0027】次に、関連単語選択処理によって得られた単語の集合から初期グラフ生成処理によって初期オントログラフを生成してグラフ作業記憶部5に格納する(ステップS17)。初期グラフは図4に示すように、ユーザが興味を持っている領域を代表する単語、この場合には「SCSI」をルート・ノードし、個々の関連単語に対応するに対して、ルート・ノードから「関連」ラベルのついたリンクを張ったものなどが考えられる。ただし、初期グラフの生成法は本発明においてここで説明した方法に限定されるものではなく、単語の品詞情報を利用して、単なる「関連」よりもより具体的なラベルをつける方法や、すでに広く一般に開示されている日常語や専門用語の言語体系を利用して、より複雑な構造を持った初期グラフを生成する方法なども考えられる。

【0028】グラフ作業記憶部5に初期グラフが格納されると、次のステップS19の共起統計量計算処理において関連単語として選択され初期グラフのノードとなっている単語の集合に対し、該単語集合から取り出したあらゆる2つの単語の組に関して同時出現についての統計量、すなわち共起統計量を計算する。共起統計量は、最も簡単には2つの単語が出現する文書の延べ数に対する該2つの単語が同時に出現する文書の延べ数の割合で定義できるが、本発明における共起統計量はここで説明したものに限定されるわけではない。他に、2つの単語が出現する段落の延べ数に対する該2つの単語が同時に出現する段落の延べ数の割合、また、2つの単語が出現する文の延べ数に対する該2つの単語が同時に出現する文の延べ数の割合なども考えられるし、これらの割合の線形結合なども考えられる。

【0029】また、共起統計量計算処理と並行して類似度計算定義を行う(ステップS21)。類似度計算処理は、関連単語として選択され初期グラフのノードとなっている単語の集合に対し、該単語集合から取り出したあらゆる2つの単語の組に関して、該2つの単語の類似度を計算する。類似度は、最も簡単には、広く一般に開示されている類語辞書を利用し、該2つの単語が類語辞書上で類語として記述されていれば類似度1を与え、そうでなければ類似度0とするという定義が考えられる。もちろん、本発明における類似度は、ここで説明した定義に限定されるものではなく、語と語の類似関係を数値化したものならばどのような定義でもよい。他に、広く一般に開示されている単語分類木を利用し、該2つの単語の分類木上での距離の逆数を類似度とすることなどが考えられる。

【0030】なお、本実施形態においては、共起統計量計算処理と類似度計算処理は並行して処理すると説明し

たが、これは、本発明において、該2つの処理が初期グラフ生成処理とグラフ変換および関係ラベル付与処理との間で行われなければならないことだけを規定するものである。すなわち、本発明において、該2つの処理は、必ずしも同時並行に行う必要はない。共起統計量計算処理を行って類似度計算処理を行ってもよいし、逆に類似度計算処理を行ってから共起統計量計算処理を行ってもよい。

【0031】共起統計量計算処理と類似度計算処理が完了すると、次にグラフ変換処理および関係ラベル付与処理を行う(ステップS23)。この処理は、グラフ作業記憶部5に記憶されたグラフに対し、共起統計量計算処理(ステップS19)により計算された共起統計量および類似度計算処理(ステップS21)により計算された類似度を利用して、グラフ変換とラベルの付与を行う処理である。

【0032】図5および図6にグラフ変換および関係ラベル付与の例を示す。まず、図6において、初期グラフにおいて単語「SCSI」から単語「カード」へは「関連」リンクが張られ、単語「SCSI」から単語「A社」へも「関連」リンクが張られている。ここで、共起統計量計算処理の結果から単語「カード」に対する単語「A社」の共起統計量が大きいことがわかると、単語「A社」は単語「カード」に包含されると判断できるため、単語「SCSI」から単語「A社」へのリンクを取り外し、単語「カード」から単語「A社」へのリンクを張るというグラフ変換を行う。また、単語「カード」と単語「A社」との関係は「A社」の品詞が固有名詞であることからインスタンスと判断し、ラベル「インスタンス」を付与する。

【0033】図6では、初期グラフにおいて単語「SCSI」から単語「高い」は「関連」リンクが張られ、単語「SCSI」から単語「安い」へも「関連」リンクが張られている。類似度計算処理の結果から単語「高い」と単語「安い」の類似度が相互に高いことが得られると、単語「高い」と単語「安い」はより結び付きが強く何らかの上位概念に包含されることがわかるので、単語「SCSI」と、単語「高い」および単語「安い」の間に新たなノードを割り込ませる形でグラフを変換する。新たに割り込ませたノードには、広く一般に開示されている単語分類木などを参照して、単語「高い」と単語「安い」を包含する単語「値段」をラベルとして付与する。単語「値段」から単語「高い」および単語「安い」へのリンクには、包含関係を示す「包含」を付与し、また広く一般に開示されている単語知識を利用して単語「値段」は物の性質を表すことを得て、単語「SCSI」から単語「値段」へのリンクにはラベルとして「性質」を付与する。

【0034】以上のようなグラフ変換および関係ラベル付与のルールをグラフ作業記憶部5に記憶されたグラフ

に繰り返し適用し、適用できるルールがなくなった場合に、グラフ変換および関係ラベル付与処理を終了する。なお、ここでは本実施形態におけるグラフ変換および関係ラベル付与のルールの例を示したが、本発明においてグラフ変換および関係ラベル付与のルールはこれらに限定されるものではない。他に、言語に関する一般的知識および広く一般に開示されている単語に関する知識を利用したルールが多く考えられる。

【0035】グラフ変換および関係ラベル付与処理を終了すると、グラフ作業記憶部5に用語辞書、すなわちオントロジをグラフとして格納したものが得られていることになる。本実施形態において生成された用語辞書を図7に示す。なお、生成された用語辞書をどのように出力するかは、本発明で規定するものではない。グラフそのままの形式でグラフィック出力装置に出力することも可能であるが、一般には、グラフ作業記憶部5に記憶された状態そのまま、あるいは単純な形式変換によって、文書理解プログラムなどの他のプログラムに渡され利用されることになる。

【0036】

【発明の効果】以上説明したように、本発明によれば、文書を構成する各単語をその位置情報とともに格納し、同一単語が含まれることに関する一次統計量を計算し、該一次統計量に基づいて関連単語を選択し、この関連単語の対象領域を表す単語のノードから各関連単語のノードに有向リンクを張ったグラフを生成し、このグラフの各2つのノードの組合せについて共起統計量を計算し、各組合せの2つの単語の類似度を計算し、共起統計量および類似度に基づいて前記グラフを変換し、リンクに関係ラベルを付与し、オントロジとして生成し、用語辞書を生成しているので、広い対象領域にわたる大量の文書に対してもコンピュータによる文書の理解、検索、分類、同合等の処理に必要とされる十分な情報を含んだ用語辞書を生成することができる。

【図面の簡単な説明】

【図1】本発明の一実施形態に係る用語辞書生成方法の処理の流れおよび該処理を実施するのに必要な構成要素の一部を示した図である。

【図2】図1の実施形態に使用されている形態素記憶部のテーブル構造を示す図である。

【図3】図1の実施形態の一次統計量計算処理で一次統計量の計算に使用される各単語の出現回数を示すテーブルである。

【図4】図1の実施形態において初期グラフ生成処理で生成された初期グラフを示す図である。

【図5】図1の実施形態においてグラフ変換および関係ラベル付与処理の例を示す説明図である。

【図6】図1の実施形態においてグラフ変換および関係ラベル付与処理の他の例を示す説明図である。

【図7】図1の実施形態において生成された用語辞書を

10

20

30

40

50

示す図である。

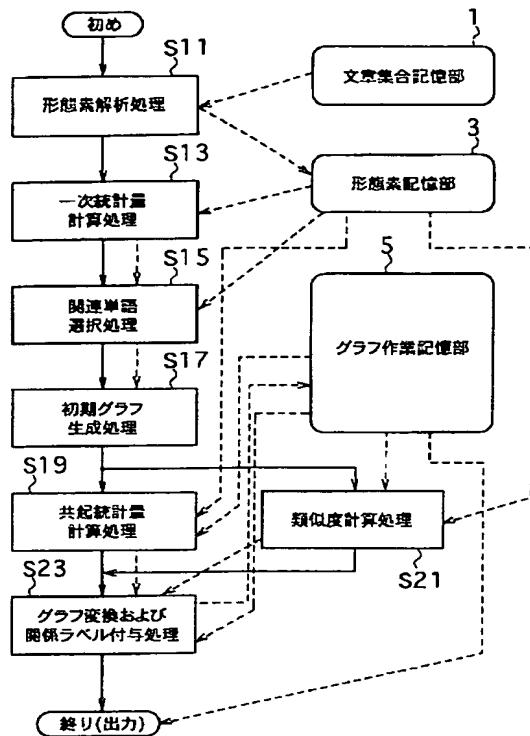
【符号の説明】

1 文書集合記憶部

3 形態素記憶部

5 グラフ作業記憶部

【図1】



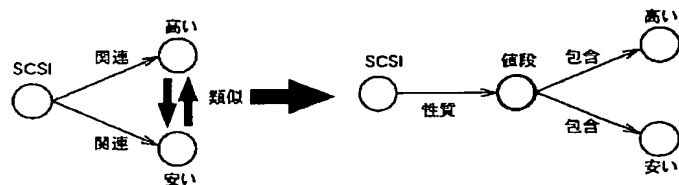
【図2】

単語	文書ID	段落ID	文ID
SCSI	0001	0001	0001
SCSI	0002	0003	0020
SCSI	0005	0001	0004
...			
カード	0001	0002	0001
...			

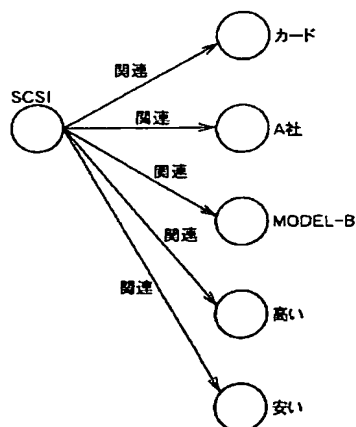
【図3】

単語	出現回数
SCSI	12345
カード	3854
...	
合計	54626

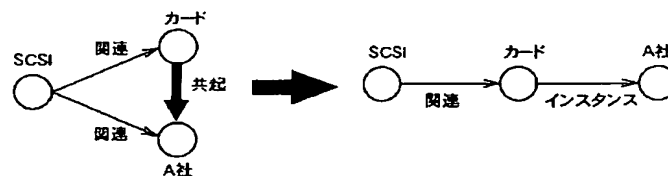
【図6】



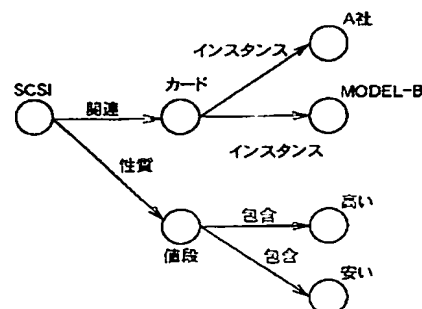
【図4】



【図5】



【図7】



(8)

特開平 1 1 - 9 6 1 7 7

フロントページの続き

(51)Int.Cl.⁶

識別記号

F I

G 0 6 F 15/403

3 2 0 D

3 3 0 C

3 5 0 C